Sohail(Neel) Sarkar

☎ 437-329-3448 ✉ sohail.sarkar@utoronto.ca 🌐 n33levo.github.io in linkedin.com/in/n33levo github.com/n33levo

Education

University of Toronto Expected May 2027
H.B.Sc. Applied Mathematics Specialist and Computer Science Minor *Toronto, ON*

Trinity College Jun 2021
B.A. Music, ACTL Classical Guitar *London, England*

Experience

Applied Scientist Intern Jun 2025 – Present
Sigma Squared, EO Ventures *Boston, MA*

- Engineered an agentic EDA to ETL for success-factors behavior prediction, reduced variance by 27%, & RMSE by 8%.
- Developed Agentic SDK with MCP wrapping, scaling 50+ workflows across data pipelines at S2 and EO Ventures.
- Co-authoring proprietary research on econometric-based performance clustering and calibration using Set and Group theory.

Data Science Intern May – Jul 2025
PSP Investments *Montréal, QC*

- Architected alpha-signal data processing pipeline answering 3,000+ analyst questions per batch.
- Reinforced post-processing with claimification-based cross-referencing; achieved 100% factual accuracy.
- Built Scala retriever layers for preprocessing & learned query routing; cut batch inference from 20 to 3 min.

MLOps Intern Jan – Apr 2025
University of Toronto *Toronto, ON*

- Delivered a distributed, hierarchical retriever enabling multi-tenant chat and scalable context routing across teams.
- Deployed RAG framework with CI/CD automaton for indexing, evaluation, and releases; cut release times by 60% (5d→2d).
- Exposed library GPU cluster with REST control-plane for LLM fine-tuning, including job scheduling and telemetry.

Software Engineering Intern Sept – Dec 2024
E.J. Pratt Library *Toronto, ON*

- Developed version control scheduling; optimized Random Forest to be 6× faster, result & 12% more shift coverage.
- Shipped OAuth-secured ticketing (MongoDB + Redis); reduced IT desk resolution time by 40% (3d→1.8d).

Projects

Talk to my PA

Ref-Rag-lite | *PyTorch, FAISS, PEFT, HotpotQA, RL (LinUCB, Thompson Sampling), MPS/CUDA, Docker* Oct 2025

- Developed Reinforcement Learning based selective expansion for token-efficient RAG, compressing most chunks to embeddings & expanding only key ones; benchmarked and achieved 48% F1, 43% EM, 50% fewer tokens, and 4× inference.

KernelForgeML | *Rust, MLIR, autotuning, CPU/GPU backends* 2025

- Developed a Rust-first ML compiler framework for transformer kernels: defined ops in MLIR, autotuned variant selection, and supported CPU + GPU backends.

Fine-Tuned Stock Analysis LLM | *Bayesian-LoRA, 4-bit quantization, PyTorch, PEFT, bitsandbytes, CUDA* Nov 2024

- Fine-tuned LLaMA-3.1-8B w 4-bit quantized Bayesian-LoRA on market reports, earnings calls & SEC filings; used Bayesian rank gating to adapt per-layer ranks while reducing compute & memory.

Recomendation Engine | *Neo4j, PyTorch, FastAPI, Dagster, AWS EKS(Kubernetes), Docker* Mar 2024

- Hybrid recommender implicit-feedback ALS on watch history & Neo4j knowledge-graph features with a trained re-ranker; Dockerized FastAPI inference & Dagster refresh pipelines and deployed on EKS w autoscaled endpoints.

Technical Skills

Languages: Python, Go, Bash, TypeScript, C/C++, Assembly, SQL
Frameworks & Libraries: PyTorch, TensorFlow, CUDA, Hugging Face, LangChain, LlamaIndex, FastAPI, FastMCP, vLLM, Triton, scikit-learn, XGBoost, OpenCV, Streamlit, Tailwind CSS, Agentic SDKs
Data & Infra: Spark, Databricks, MLflow, AWS Glue, Athena, Presto, Hive, Kafka, Docker, k8s, AWS, REST, GraphQL, Jenkins
Databases: Neo4j, Pinecone, PostgreSQL, MongoDB, Redis, SQLite